ЗП

УДК 004.853

О.М. Почанский
Харьковский национальный университет радиоэлектроники, Украина
Украина, 61166, г. Харьков, пр. Ленина 14, *pochansky.oleg@yandex.ru*

Социальное индексирование Web-документов для семантического поиска

O.M. Pochanskiy
Kharkiv National University of Radioelectronics, Kharkiv
Ukraine, 61166, Kharkiv, Lenina st., 14

Social Indexing of Web-Documents for Semantic Search

О.М. Почанський
Харківський національний університет радіоелектроніки, м. Харків, Україна
Україна, 61166, м. Харків, пр. Леніна, 14

Соціальне індексування Web-документів для їх семантичного пошуку

В данной работе вводится понятие нового термина критерия поиска информации «социальный индекс», главной задачей которого является вычисление значимости любого Web-документа для конечного пользователя в зависимости от его текущих интересов. Методика его определения и применения при создании независимых поисковых систем рассмотрены в данной статье.
Ключевые слова: социальный индекс, семантический поиск, база знаний, кластерный анализ, плагин.

Social index is a new standard of information search, which is defined in the given work. Its main task is to calculate the significance of any Web-document to the end user depending on his/her current task. The method of its determination and application for the creation of independent search systems is considered in the text of this article.
Key words: social index, semantic search, knowledge base, claster analysis, plug-in.

У даній роботі вводиться поняття нового терміна критерію пошуку інформації – соціальний індекс. Його головним завданням є – обчислення важливості будь-якого Web-документа для кінцевого користувача в залежності від його поточних інтересів. Методика визначення та застосування соціального індексу при створенні незалежних пошукових систем розглянуті в тексті даної статті.
Ключові слова: соціальний індекс, семантичний пошук, база знаній, кластерний аналіз, плагін

Введение

Всемирная сеть предоставляет каждому человеку возможность в самореализации, вовлекая тем самым все больше новых пользователей. А с возникновением социальных сетей появилась возможность не только постоянно общаться в режиме «on-line», но и обмениваться разного рода информацией и данными. При этом социальным сетям с их техническими и научными возможностями не уделяется должного внимания.

Цель работы. Основной целью данной статьи является введение нового критерия поиска по социальной направленности искомой информации. Под социальной направленностью понимается предрасположенность определенной категории людей к интересующей их информации. В данной работе предложенный критерий будет называться социальным индексом.

Постановка задачи. Предлагается разработать интеллектуальную систему извлечения знаний из сети Internet, которая будет предлагать пользователю Web-документы по искомой им тематике, на основе их популярности у других пользователей, со схожими с ним интересами. При этом популярным будет считаться тот электронный документ, на который подписалось (выбрали как основной источник информации) наибольшее количество пользователей.

Анализ последних исследований и публикаций

В рамках исследуемой предметной области можно выделить следующие работы, посвященные поиску информации на основе построения социальных связей между различными пользователями сети Internet:

1. On designing and implementing a collaborative system using the distributed-object model of Java RMI [1].
2. Profiling and matchmaking strategies in support of opportunistic collaboration [2].
3. Recommending collaboration with social networks: A comparative evaluation [3].
4. Social networks and Social information filtering on Digg [4].

В первой работе проблема поиска пользователями интересующих их данных решается путем создания общего хранилища информации. При этом каждый пользователь данной системы формирует свои ограничения (фильтры) на искомые им документы. Тем самым определяется список свойств, характеризующий искомую информацию, который в свою очередь может быть расширен за счет фильтров, заданных другими пользователями. В результате точность классификации любого документа, хранимого в данной системе, повышается.

Основные достоинства данной системы:

1. Организация единого независимого хранилища документов по искомым пользователям тематикам.
2. Улучшенная точность классификации документов в рамках сформированного хранилища.
3. Использование фильтров, созданных различными пользователями в рамках единой среды.

Недостатки данной системы:

1. Отсутствие критерия проверки корректности задания пользователем фильтра для поиска интересующих его документов.
2. Чувствительность системы к программному обеспечению компьютера пользователя.
3. Размер хранилища ограничен документами пользователей, которые зарегистрированы в системе.
4. Система работает только с зарегистрированными пользователями.
5. Эффективность системы зависит от общего количества пользователей, их честности и объективности.

Во второй работе рассмотрена проблема организации обмена информацией между различными пользователями со схожими интересами в любых предметных областях. Это задача решается путем создания мультиагентной системы.

С помощью данных агентов, для каждого пользователя формируется индивидуальный профиль, в котором хранится информация, заполненная им, а также те данные, которые были получены в процессе анализа его активности во время работы с данной системой. Для этого каждый пользователь должен создать свою рабочую среду (в рамках одной исследуемой им предметной области), в которой он указывает свои пер-

сональные данные (e-mail, ФИО, контактный телефон, свои интересы и т.д.), рабочие проекты и сопутствующие документы или ссылки на Web-ресурсы. Эта информация будет составлять основу его индивидуального профиля. Затем данная система проанализирует полученные данные и выделит список основных ключевых слов, которые встречаются в текстовых ресурсах, указанных пользователем. В дальнейшем это позволяет находить новые документы, с которыми работали другие пользователи, на основании совпадения основных ключевых слов, характерных для активной в текущий момент рабочей среды. Также в процессе работы пользователя с различными документами в рамках одной активной рабочей среды формируется рейтинг наиболее активно используемых источников информации (документов, Web-ресурсов). Исходя из этого, составляется список наиболее популярных источников для каждого пользователя в рамках одной рабочей среды.

Обмен данными между пользователями осуществляется на основании схожести предметной области их рабочих сред. Он выполняется путем равносторонней отсылки документов или ссылок на Web-страницы, которые отсутствуют у одного из пользователей. В дополнение к обмену информацией системой предусмотрены основы коммуникации по средствам текстовых сообщений между пользователями со схожей предметной областью исследований в рамках активной рабочей среды.

Основным достоинства данной системы:

1. Механизм обмена данными между пользователями.
2. Возможность коммуникации пользователей с общими интересами.
3. Автоматическое обновление профиля пользователя в процессе работы в рассматриваемой системе.
4. Возможность пользователя создавать несколько рабочих сред для каждой предметной области.
5. Организацию рейтинга популярности документов для каждого пользователя в рамках активной рабочей среды.

Основные недостатки данной системы:

1. Система ориентирована на закрытые социальные группы в рамках одной организации с ограниченным документооборотом.
2. Требует установки специального программного обеспечения, которое нуждается в предварительной настройке.
3. Одновременно пользователь может работать только в одной рабочей среде.
4. Рабочие среды, в которых работает пользователь, никак не связаны между собой.
5. Система не хранит полнотекстовые копии Web-ресурсов, которые в своей рабочей среде отметили пользователи. В результате возможно появление ссылки на не существующий ресурс.

В третьей работе рассматривается проблема обмена информацией между пользователями, которые объединены в социальную сеть в рамках одной организации. Для этого проектируется соответствующая система, в которой путем опроса пользователей оценивается эффективность установления ограничений на обмен информацией между различными организационными структурами в рамках одной организации. Таким образом, было выявлено, что для самих пользователей приоритетным является получение необходимой информацией в сжатые сроки от источника, которому он может доверять в независимости от его месторасположения в иерархии системы. Сам же механизм поиска авторитетного ресурса должен основываться на наличие доступных ресурсов и опыта разработчиков системы.

Основные достоинства данной системы:

1. Быстрый обмен знаниями между пользователями по интересующим их вопросам в рамках одной организации.

2. Наличие возможности коммуникации между пользователями с общими интересами.

3. Возможность решать трудные задачи коллективной работой пользователей.

Основные недостатки системы:

1. Система ориентирована на закрытые социальные группы в рамках одной организации с ограниченным документооборотом.

2. Слабо развит документооборот между пользователями.

3. В системе не предусмотрено общеё хранилище документов.

4. При выполнении коллективных задач затруднена оценка степени участия каждого отдельного пользователя.

5. Отсутствует механизм поиска информации по ключевым словам внутри системы.

В четвертой работе рассматривается проблема выявление значимых данных для конечного пользователя на базе информационного ресурса Digg [5].
Данный ресурс представляет собой динамическую Web-страницу, на которой каждый зарегистрированный пользователь может оставить ссылку. Также он может проголосовать за другие источники информации, добавленные остальными пользователями системы. При этом любая новость на ресурсе Digg может быть прокомментирована зарегистрированным пользователем. В результате на первой Web-странице данного ресурса отображаются источники информации с наибольшим рейтингом. Для персонализации данной странички разработчики системы предлагают пользователям объединяться в социальные подгруппы, в рамках которых рейтинг отмеченных ими ссылок выше, чем остальных. Таким образом, осуществляется попытка уменьшения эффекта влияния на рейтинг наиболее активных пользователей системы, которые часто его неоправданно завышают для некоторых источников информации.

Основные достоинства данной системы:

1. Постоянное обновление ресурса за счет добавления новых источников различными пользователями.

2. Индивидуальный расчет значимости источников информации за счет объединения пользователей в социальные группы.

3. Реализация концепции Web 2.0 (каждый пользователь может в режиме реального времени дополнять Web-ресурс новой информацией).

Основные недостатки системы:

1. Отсутствует механизм деления пользователей по интересам, который мог бы улучшить объективность рейтинга источников информации.

2. В системе нет функции пополнения данных за счет интеграции с другими ресурсами.

3. Ограниченность системы. Она ориентирована на работу только с новостными ресурсами.

4 Система поиска информации на основе социального индексирования

4.1 Конструктивные особенности

Исходя из проделанного анализа основных наработок выполненных в данной предметной области было принято решение: разработать специализированную поисковую систему, которая должна обладать следующим конструктивными особенностями:

1. Учитывать основные интересы пользователя при поиске нужной ему информации.

2. Сортировать найденную информацию в соответствии с текущими интересами пользователя.

3. Оценивать привлекательность Web-ресурса для конечного пользователя.

4. Работать с различными ресурсами сети Internet, доступ к которым имеет любой пользователь.

5. Реализовывать основные преимущества концепции Web 2.0 (возможность оценивать и комментировать прочитанные источники информации) и Web 3.0 (поиск источников информации на основе их семантики).

6. Объединять пользователей из различных социальных сетей на основании общих тем.

7. Использовать профиль пользователя в социальных сетях для автоматического заполнения регистрационной формы.

Таким образом, для выполнения описанных выше конструктивных особенностей было принято решение о создании системы, способной выполнять поиск информации, основываясь на собственном критерии оценки данных – социальном индексе. Он будет определять степень важности информации в зависимости от текущих интересов пользователя.

4.2 Алгоритм работы

Общий алгоритм работы системы поиска информации на основе социального индекса представлен следующей блок-схемой (рис. 1).

Данная структурная схема выступает в роли опорного плана взаимодействия пользователя и системы, основанной на методе поиска информации по социальному индексу. В ее основу включены основные функциональные элементы, которые посвящены главным образом регистрации нового пользователя в системе. На этом этапе, помимо указания личных данных о себе, пользователь может подключить Rdf-файл своего социального профиля (при его наличии), составленного в соответствии со стандартом Semantic Web FOAF [6]. Также существует возможность установки специально разработанного плагина. Он выполняет две основных функции:

1. Позволяет пользователю отслеживать все актуальные изменения в системе (доступ к интересующей его информации по «клику»).

2. Позволяет системе постоянно обновлять свою базу знаний путем мониторинга активности пользователя сети Internet (плагин передает адреса Web-страниц, которые посещались, из которых считывается название и список основных ключевых слов, взятых из meta-данных).

После активации учетной записи пользователь переходит на специально созданную для него Web-страницу, где он может просмотреть список наиболее популярных источников среди пользователей со схожими интересами. Результат представляется в виде динамической таблицы, в которой указывается тематика и перечень востребованных источников с указанием их описания и ссылкой для перехода. Также существует возможность поиска нужной информации по всей базе знаний системы с указанием ключевых слов. В этом случае схожесть интересов различных пользователей не учитывается.

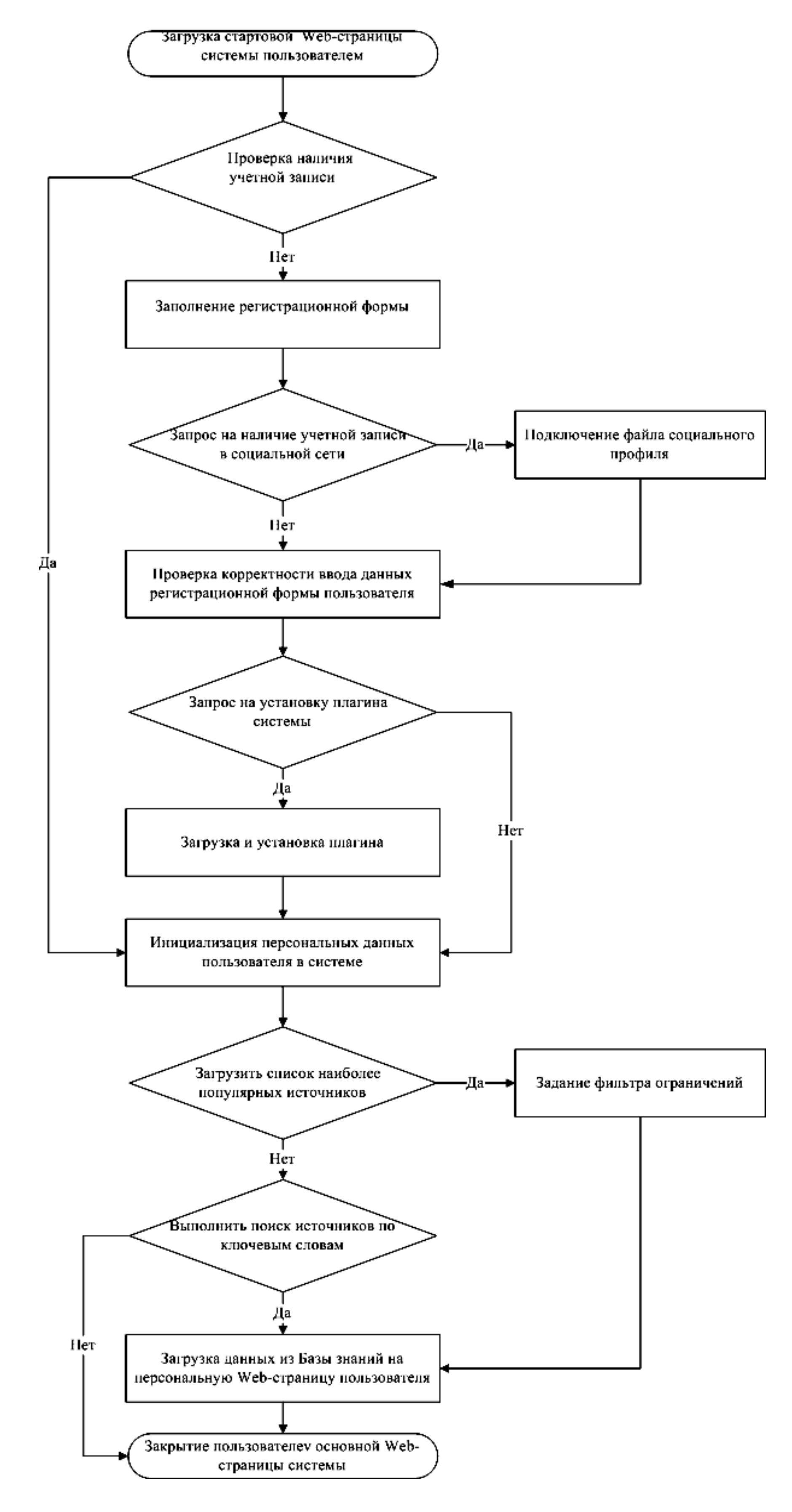


Рисунок 1 – Общий алгоритм проектируемой системы

4.3 База знаний

Роль основного хранилища данных в разрабатываемой системе поиска информации на основе социального индекса выполняет база знаний. Она представлена онтологией, которая хранится в виде OWL-файла. Основные принципы её органи-

зации для создания базы знаний интеллектуальных систем изложены в статье, посвященной соответствующей тематике [7]. В базовой версии она состоит из трех основных терминов: User (Пользователь), Interests (Интересы пользователя) и WebSource (Название страниц, которые посещал пользователь), взаимодействие которых показаны на рис. 2.

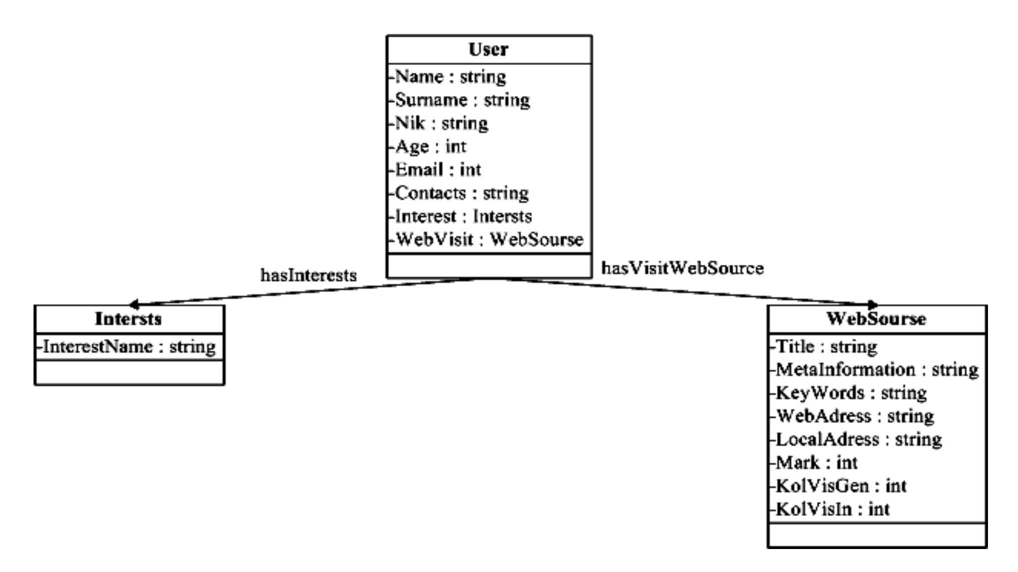


Рисунок 2 – Основные термины базовой версии онтологии рассматриваемой системы

Главный принцип работы рассматриваемой базы знаний заключен в следующем:

1. Данные, полученные после регистрации пользователя, поступают в термины «User» и «Interests». При этом в первом хранится общая информация о пользователе (ФИО, возраст, e-mail и т.д.), а во втором – содержится наименование основных его интересов (выбранных из предложенного списка при регистрации). Эти термины связаны между собой свойством hasInterests, с помощью которого элементы Interests являются частью User.

2. В термин «WebSource» данные поступают из плагина, который закачивается пользователем при регистрации. Они представляют собой ссылки на посещяемые им Web-страницы, а также информацию, полученную при анализе их HTML-кода. WebSourse также связан с термином «User» свойством hasVisitWebSourse по аналогии с термином Interests.

3. После накопления данных, полученных от пользователей системы, они могут быть выведены из базы знаний в соответствии со сформированным запросом.

Если основная функция базы знаний в системе поиска информации на основе социального индексирования заключается в основном только в хранении/выводе информации, то с основной задачей определения степени значимости для каждого пользователя справляется метод социальной оценки Web-документов, описанный ниже.

4.4 Метод социальной оценки Web-документов

Данный метод позволяет вычислить оценку значимости (социальный индекс SI) Web-документа для каждого пользователя системы поиска информации на основе социального индексирования. Для этого используются значения соотношений общего количества к количеству посещений пользователей с определенной группой интересов CU с диапазоном значений [0..1] при условии $CU \in R$. Также при расчете социального индекса учитывается средняя оценка привлекательности Web-документа MU с диапа-

зоном значений [0..10] при условии $MU \in Z$ среди пользователей со схожими интересами. Данная оценка выставляется с помощью использования установленного плагина системы.

В результате оценка значимости SI любого источника, информация будет состоять из двух независимых друг от друга параметров. Но для того, чтобы вычислять социальный индекс для каждого Web-документа максимально объективно, необходимо задать значения некого шаблона, у которого значения CU и MU максимальны и равны 1 и 10 соответственно. В этом случае, чем ближе к нему источник информации, тем выше у него SI. Исходя из этого, взяв за основу запись Евклидового расстояния для сравнения объектов с двумя независимыми параметрами [8], социальный индекс будет вычисляться по следующей формуле:

$$SI_i = \sqrt{(MU_t - MU_i)^2 + (CU_t - CU_i)^2}, \qquad (1)$$

где SI_i – социальный индекс i-го Web-документа;

MU_t – оценка значимости шаблона (всегда максимальна);

MU_i – средняя оценка значимости i-го Web-документа среди пользователей со схожими интересами;

CU_t – соотношения общего количества посещений шаблона к количеству посещений пользователей с определенной группой интересов (всегда максимально);

CU_i – соотношения общего количества посещений i-го Web-документа к количеству посещений пользователей с определенной группой интересов (всегда максимально).

Основываясь на полученной формуле социального индекса, основные действия работы метода социальной оценки Web-документа буду состоять из следующих шагов:

1. Получение списка текущих значений MU_i и CU_i для i-го Web-документа (из базы знаний), поставленных пользователями с различными интересами.

2. Выбор значений параметров MU_i и CU_i, установленных для пользователей с определенным типом интересов.

3. Расчет социального индекса для i-го Web-документа согласно формуле (1).

4. В случае открытия i-го Web-документа пересчитать значение CU_i по всему списку значений с учетом списка интересов пользователя, загрузившего его, и записать его в базу знаний системы.

5. В случае оценки i-го Web-документа пересчитать значение MU_i по всему списку значений с учетом списка интересов пользователя, отметившего его, и записать его в базу знаний системы.

Таким образом, по запросу пользователя к системе поиска информации по социальному индексу, будет выведен список Web-документов, в порядке их соответствия его текущим интересам.

Выводы

В данной работе предлагается новый критерий поиска информации – социальный индекс. Он позволяет оценивать значимость Web-документа в зависимости от текущих интересов каждого пользователя. Таким образом, ресурсы сети Internet могут быть проиндексированы не только по ключевым словам, но и по принадлежности их к определенной социальной группе людей, к примеру, фанатам спорта (футбола). В результате формируется новая социальная сеть, в которой связаны между собой пользователь, его интересы в различных предметных областях и Web-документы, соответст-

вующие этим интересам. Эта особенность выгодно отличает предложенную систему поиска информации на основе социального индексирования от схожих решений других авторов, рассмотренных выше в статье.

В дальнейшем планируется практическая разработка предложенной в статье системы с последующим сравнением эффективности её работы с другими аналогами, а также улучшение механизма интеграции социального профиля пользователя в базу знаний предлагаемого решения.

Литература

1. Raje R.R. On On designing and implementing a collaborative system using the distributed-object model of Java RMI / R.R. Raje, S. Mukhopadnyay, M. Boyles and others // Progress in computer research – Nova Science Publishers, Inc. Commack. – NY, USA, 2001. – P. 123-134.
2. Vivacqua A. Profiling and matchmaking strategies in support of opportunistic collaboration / A. Vivacqua, M. Moreno, J. Souz // Lecture notes in computer science – Springer-Verlag, Berlin Heidelberg, 2003. – P. 162-177.
3. McDonald D.W. Recommending collaboration with social networks: A comparative evaluation / D.W. McDonald // Proceedings of the SIGNCHI conference on Human factors in computing systems. – 2003. – P. 593-600.
4. Lerman K. Social networks and Social information filtering on Digg / K. Lerman // Computing Research Repository – CORR. – 2006. – 8 p.
5. Режим доступа : http://digg.com/
6. Режим доступа : http://foaf-project.org/
7. Вороной О.С. Засоби інтеграції онтологій предметних областей для створення баз знань інтелектуальних навчальних систем / О.С. Вороний, Г.А. Єгошина // Искусственный интеллект – 2010. – № 2. – С. 124-130.
8. Мандель И.Д. Кластерный анализ / И.Д. Мандель. – М : Финансы и статистика, 1988. – 176 с.

Literatura

1. Raje R.R. Progress in computer research. Nova Science Publishers, Inc. Commack. NY, USA. 2001. P. 123-134.
2. Vivacqua A. Lecture notes in computer science. Springer-Verlag. Berlin Heidelberg. 2003. P. 162-177.
3. McDonald D. W. Proceedings of the SIGNCHI conference on Human factors in computing systems. 2003 P. 593-600.
4. Lerman K. Computing Research Repository. CORR. 2006. 8 p.
5. http://digg.com/
6. http://foaf-project.org/
7. Voronoj O.S. Iskusstvennyj intellect. 2010. № 2. S. 124-130.
8. Mandel' I.D. Klasternyj analiz. M: Finansy i statistika. 1988. 176 s.

O.M. Pochanskiy

Social Indexing of Web-Documents for Semantic Search

This article describes the model of the system search, which is based on popularity of any information sources among users with different interests. This system applies a specially designed measure, i.e. a social index. Its main task declares the numerical equivalent of significance degree of any Web-document among users with different interests. This value is calculated using a specially designed formula given in the text.

The proposed system scheme consists of two parts.

The first part consists of the Web-service program, which allows registered users to search information by key words or using social index. The results of the first one can be sorted according to their degree of importance to users with similar interests.

The second part offers the user to install a plugin during the process of the registration, which automatically builds into his browser. It allows users to get in real-time mode the latest information from the Web-service system (for example, the most popular sites among another users who have common interests with him/her), and to evaluate any of the Web-documents on the Internet, which he/she read. Another function of the plugin is to send information to the system server with permissions of the user about the Web-page (name, address and etc.), which he/she visited. This information is used to calculate the social index and update the system knowledge base.

The registration of a new user in the system can be performed automatically if his/her account in any of the social networks compiled according to FOAF (W3C standard).

As a result, the final model of the system can index the Internet resources, not only by keywords but also by their belonging to a particular social group of people, for example, fans of sports (football). Thus, a new social network is created, in which all users, their interests in different fields and Web-documents relate to each other.

Статья поступила в редакцию 02.12.2011.